TEXAS A&M UNIVERSITY

COLLEGE STATION, TEXAS 77843-3143

AD A117113

AUTOREGRESSIVE SPECTRAL ESTIMATION

AND FUNCTIONAL INFERENCE

by Emanuel Parzen

Institute of Statistics

Texas A&M University

Technical Report No. N-29

June 1982

DTIC SELECTED JUL 21 1982

DTIC FILE COPY

07 19 012

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER TR - N-29	2. GOVT ACCESSION NO. AD-A117 113	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) AUTOREGRESSIVE SPECTRAL ESTIMATION AND FUNCTIONAL INFERENCE		5. TYPE OF REPORT & PERIOD COVERED Technical
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Emanuel Parzen		8. CONTRACT OR GRANT NUMBER(s) ONR N00014-82-MP-2001 ARU DAAG 29-80-C-0070
9. PERFORMING ORGANIZATION NAME AND ADDRESS Texas A&M University Institute of Statistics College Station, TX 77843		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE June 1982
		13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS(If different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Time series analysis, model identification, spectral estimation, ARMA model, information divergence, functional inference.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Functions used to describe the probability distributions of time series (both Gaussian and non-Gaussian) are introduced (section 1). The concept of type of a time series is defined (section 2). Autoregressive spectral densities are defined (section 3). Order determining criteria are motivated (section 5) through the concept of model identification by estimating information. An approach to empirical spectral analysis is suggested (section 4).

AUTOREGRESSIVE SPECTRAL ESTIMATION AND
FUNCTIONAL INFERENCE

Emanuel Parzen
Institute of Statistics, Texas A&M
University, College Station, TX 77843 USA

The theory of statistical spectral analysis has a long tradition of applications to the sciences of oceanography, underwater sound, anti-submarine warfare, and signal processing. The probability theory of spectral analysis was pioneered by Norbert Wiener (1930) under the stimulus of the question of how to describe mathematically the randomly changing surface waves on the Charles River on which he daily gazed from his office window at M.I.T. The statistical theory of spectral analysis began to be developed in the 1950's, and was early applied to the ocean sciences by Walter Munk (1957) who was in close touch with relevant research by statisticians such as Blackman and Tukey (1959), Goodman (1957), and Parzen (1957, 1958, 1961). (Parzen's research was stimulated by his work at Hudson Laboratories of Columbia University, an ONR sponsored project on underwater sound). Carter and Nuttall (1981) describe the interests of the navies of the world in signal processing systems to detect, classify, and localize the sources of signals.

A discussion of the naval applications of time series analysis requires attention to the unique characteristics of signals and noises encountered in the naval environment, a discussion of quantization and time/frequency sampling techniques. These topics are not considered in this paper which aims to discuss abstractly the statistial analysis of a discrete parameter time series $Y(t)$, $t=0, \pm 1, \ldots$. We aim to make the ocean scientist aware of new ways of thinking about statistical problems that are provided by learning how to formulate statistical inference problems as _functional inference_ problems that can be approached using techniques inspired by the pioneering research on spectral density estimation, especially research on autoregressive spectral estimation and maximum entropy spectral estimation. I use "functional inference" to denote a special case of "abstract inference" as defined by Grenander (1981).

To describe what is meant by functional inference, note that many statistical problems are concerned with the estimation of the parameters of probability models (models for the probability distributions of observed samples). The parameters are usually defined to be a finite dimensional vector $\theta = (\theta_1, \ldots, \theta_k)$. Functional inference defines the parameters to be functions [such as $f(\omega)$, $0 < \omega \le 1$] on the unit interval, or the unit square, or the unit hyper-cube in higher dimensions.

Functions used to describe the probability distributions of time series (both Gaussian and non-Gaussian) are introduced (section 1). The concept of type of a time series is defined (section 2). Autoregressive spectral densities are defined (section 3). Order determining criteria are motivated (section 5) through the concept of model identification by estimating information. An approach to empirical spectral analysis is suggested (section 4).

1. Describing Gaussian and Non-Gaussian Stationary Time Series

To say that a function $Y(t)$, $t=0, \pm 1, \ldots$ is a time series (or a stochastic process or a random function) is to say that the function $Y(t)$ does not have at each time t a definite value but has an ensemble of values which it assumes in

accordance with a probability distribution. To describe the probability distri-
bution of a continuous random variable $Y(t)$ one uses one (or more) of the
following functions of $-\infty < y < \infty$ and $0 \leq u \leq 1$ [see Parzen (1979)]:

distribution function $\qquad F_{Y(t)}(y) = P_r[Y(t) \leq y]$,

probability density function $\qquad f_{Y(t)}(y) = \dfrac{d}{dy} F_{Y(t)}(y)$

quantile function $\qquad Q_{Y(t)}(u) = F_{Y(t)}^{-1}(u)$

quantile density function $\qquad q_{Y(t)}(u) = \dfrac{d}{du} Q_{Y(t)}(u)$

density-quantile function $\qquad fQ_{Y(t)}(u) = \{q_{Y(t)}(u)\}^{-1}$

$\qquad\qquad\qquad\qquad\qquad\qquad = f_{Y(t)} \{Q_{Y(t)}(u)\}$

Moments are described by

mean value function $\qquad \mu(t) = E[Y(t)]$,
Covariance kernel $\qquad K(s,t) = \text{Cov}[Y(s), Y(t)]$.
standard deviation function $\qquad \sigma(t) = \{K(t,t)\}^{\frac{1}{2}}$
A time series is called <u>Gaussian</u> if any linear combination $\sum_{i=1}^{m} c_i Y(t_i)$ is

normally distributed. The standard Gaussian distribution and density functions
are denoted

(1) $\qquad \phi(y) = \int_{-\infty}^{y} \phi(y)\, dy, \quad \phi(y) = \dfrac{1}{\sqrt{2\pi}}\, e^{-(\frac{1}{2})y^2}$

The standard Gaussian quantile function is denoted $\phi^{-1}(u)$. Then

$\qquad F_{Y(t)}(y) = \phi\left(\dfrac{y - \mu(t)}{\sigma(t)}\right)$

(2) $\qquad Q_{Y(t)}(u) = \mu(t) + \sigma(t)\, \phi^{-1}(u)$.

The time series is called <u>covariance stationary</u> if $K(s,t)$ is a function only
of $|s-t|$. We then define the <u>covariance function</u>.
(3) $\qquad R(v) = K(t, t+v), \quad v = 0, \pm 1, \ldots$.
It is a <u>non-negative definite</u> function, in the sense that

(4) $\qquad \sum_{i,j=1}^{n} c_i c_j R(v_i - v_j) \geq 0$

for any integer n, any complex numbers $c_1, \ldots, c_n$, and lags $v_1, \ldots, v_n$. We use c^*
to denote the complex conjugate of c.

When $\sum_{v=-\infty}^{\infty} |R(v)| < \infty$, we define the <u>power spectrum</u> by

(5) $\qquad S(\omega) = \sum_{v=-\infty}^{\infty} e^{-2\pi i v\omega} R(v), \quad 0 \leq \omega \leq 1$.

The frequency variable ω is usually assumed to vary in the interval $-0.5 \leq \omega \leq 0.5$,
but only the interval $0 \leq \omega \leq 0.5$ has physical signficance. We prefer to take $0 \leq \omega \leq 1$
as the domain of ω, because this choice is more convenient for developing iso-
morphisms between spectral analysis and non-parametric data modeling using
quantile and density-quantile functions [introduced in Parzen (1979)]. The
spectral representation of $R(v)$ is

(6) $\qquad R(v) = \int_{0}^{1} e^{2\pi i v\omega} S(\omega)\, d\omega$.

In my judgement the results of a statistical analysis of a time series are
most insightful when expressed in units of $R(0)$, the variance. Thus the
covariance function $R(v)$ is replaced by the correlation function
(7) $\qquad \rho(v) = \dfrac{R(v)}{R(0)} = \text{Corr}[Y(t), Y(t+v)]$.

The probability law of a Gaussian time series is completely determined by the mean value function and the covariance kernel. The correlation structure of the time series is described by the correlation function $\rho(v)$ and various functions (spectral density, and autoregressive-moving average schemes) equivalent to $\rho(\cdot)$.

For non-Gaussian time series the dimensionless dependence between pairs of variables $Y(t)$ and $Y(t+v)$ is described by a sequence of dependence density functions $d_v(u_1,u_2)$ defined as follows: for $0 \leq u_1, u_2 \leq 1$

$$(8) \qquad D_v(u_1,u_2) = F_{Y(t), Y(t+v)}(Q_{Y(t)}(u_1), Q_{Y(t+v)}(u_2))$$

$$(9) \qquad d_v(u_1,u_2) = \frac{\partial^2}{\partial u_1 \partial u_2} D_v(u_1,u_2).$$

One can show that in the normal case

$$(10) \qquad \rho(v) = \int_0^1 \int_0^1 \phi^{-1}(u_1)\, \phi^{-1}(u_2)\, dD_v(u_1,u_2)$$

Other measures of dependence that can be considered include

$$(11) \qquad \rho U(v) = \frac{1}{12} \int_0^1 \int_0^1 (u_1-0.5)\,(u_2-0.5)\, dD_v(u_1,u_2)$$

which is the correlation function of the transformed time series

$$(12) \qquad U(t) = F_{Y(t)}(Y(t)),$$

called the rank-transform of the original time series. To form $U(t)$ one appears to need to know the distribution function of $Y(t)$; in the stationary case, one can estimate it, and therefore $U(t)$, from the sample. The theory and application of $U(t)$ requires further development. But in my judgement it provides a basis for a functional inference point of view to develop methods of time series analysis that are not based entirely on the Gaussian model.

In this paper we discuss methods of time series analysis based on the conventional correlation function $\rho(v)$. It is a non-negative definite function satisfying $\rho(0) = 1$. Therefore there exists a function $F(\omega)$, $0 \leq \omega \leq 1$, called the spectral distribution function such that

$$(13) \qquad \rho(v) = \int_0^1 e^{2\pi i v\omega}\, dF(\omega)$$

If $\rho(v)$ is summable, then the derivative $f(\omega) = F'(\omega)$ can be explicitly calculated by

$$(14) \qquad f(\omega) = \sum_{v=-\infty}^{\infty} e^{-2\pi i v\omega} \rho(v), \qquad 0 \leq \omega \leq 1.$$

We call $f(\omega)$ the spectral density function.

Note that $F(0) = 0$, $F(1) = 1$, and

$$(15) \qquad F(\omega) = \int_0^\omega f(\omega')\, d\omega', \qquad 0 \leq \omega \leq 1$$

Also $F(0.5) = 0.5$ since $f(\omega)$ is an even function in the sense that $f(1-\omega)=f(\omega)$. When analyzing data, one computes a modified spectral distribution function $F_+(\omega)$ which is defined for $0 \leq \omega \leq 0.5$ by $F_+(\omega) = 2\,F(\omega)$.

For a spectral density $f(\omega)$ obeying suitable conditions, one can define the inverse-correlation function [see Cleveland (1972), Parzen (1974), Chatfield (1979)]

$$(16) \qquad \rho i(v) = \int_0^1 e^{2\pi i v\omega}\, f^{-1}(\omega)\, d\omega \div \int_0^1 f^{-1}(\omega)\, d\omega$$

and the cepstral-correlation function [see Wahba (1980) for an application]

$$(17) \qquad \gamma(v) = \int_0^1 e^{2\pi i v\omega}\, \log f(\omega)\, d\omega$$

It should be noted that the inverse-correlation function is non-negative definite. However the cepstral-correlation function is not. These new types of correlation functions are introduced because they may provide more parsimonious parametrizations in the sense that they decay to 0 faster than does the correlation function. Statistical inference (from a sample) of the probability
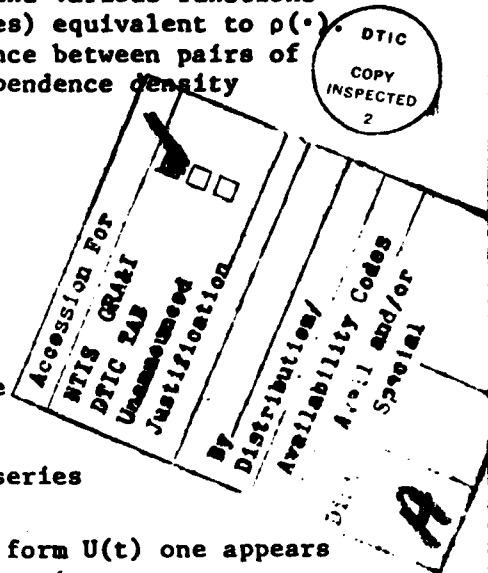
law of a time series often achieves greatest statistical efficiency by using the most parsimonious parametrizations. Thus to form estimators $\hat{f}(\omega)$ of the spectral density $f(\omega)$ from a raw estimator $\tilde{f}(\omega)$, greater precision may be attained by first forming estimators $\{f^{-1}(\omega)\}^{\wedge}$ and $\{\log f(\omega)\}^{\wedge}$ of the inverse or logarithm of the spectral density. Autoregressive spectral estimation may be regarded as an approach to estimating $f(\omega)$ by first estimating $f^{-1}(\omega)$ [Durrani & Arslanian(1982)].

2. Memory Type of a Time Series

One approach to the comparative behavior of spectral estimation methods is an empirical one which compares estimators computed for various examples of time series; see Beamish and Priestley (1981) and Kay and Marple (1981). To draw general conclusions from such studies it is necessary to define types of time series. The notion of memory is introduced in Parzen (1982) to define types.

Memory refers to how far in the future (after a specified time t_0) one can effectively predict the time series given observations on its infinite past (up to t_0). The infinite memory one step ahead predictor of a Gaussian zero mean time series is denoted $Y^\mu(t) = E[Y(t)|Y(t-1), Y(t-2),...]$, with prediction error, or innovation, $Y^\nu(t) = Y(t) - Y^\mu(t)$. The innovations have a basic property of being white noise, and are also denoted $\varepsilon(t)$. The mean square prediction error, measured in units of the variance $R(0)$ of $Y(t)$, is denoted $\sigma_\infty^2 = E[|Y^\nu(t)|^2] \div E[|Y(t)|^2]$, and satisfies

$$(1) \qquad \log \sigma_\infty^2 = \int_0^1 \log f(\omega) \, d\omega$$

when $Y(\cdot)$ is a <u>stationary Gaussian zero mean</u> time series. We classify such time series as follows:

(1) a <u>no-memory</u>, or white noise, time series if it satisfies either of the equivalent conditions: $\rho(v) = 0$ for $v>0$; $f(\omega) = 1$, $0<\omega\leq1$; $\sigma_\infty^2 = 1$.

(2) a <u>short memory</u> time series if its correlation function $\rho(v)$ is summable, and its spectral density function $f(\omega)$ is bounded above and below in the sense that the dynamic range of $f(\omega)$

$$(2) \qquad DR(f) = \max_{0<\omega\leq1} f(\omega) \div \min_{0<\omega\leq1} f(\omega)$$

satisfies $1<DR(f) < \infty$. Then $0<\sigma_\infty^2<1$.

(3) a long memory time series if it is neither no memory nor short memory, and $\sigma_\infty^2 = 0$. Alternatively a long memory time series is one which is non-stationary or non-ergodic. It usually has components representing cycles or trends. An example of a long memory time series is

$$(3) \qquad Y(t) = A \cos 2\pi\omega_0 t + B \sin 2\pi\omega_0 t,$$

where A and B are independent $N(0,\sigma^2)$ random variables; its correlation function $\rho(v) = \cos 2\pi\omega_0 v$ is not summable, and $\sigma_\infty^2 = 0$. A band limited time series has a correlation function that decays to 0 as v tends to ∞ at the rate of $1/v$; it has long memory.

An important example of a long memory time series arises when $Y(t) = S(t) + N(t)$, where the signal $S(t)$ is long memory and the noise $N(t)$ is short memory or no memory. Kay and Marple (1981), p. 1408 consider a signal which is the sum of three sinusoids at frequencies .10, .20, and .21 respectively, and signal to noise ratios (SNR) of 10, 30, and 30 dB respectively where SNR is defined as the ratio of the sinusoid power to the total power in the noise process.

Our approach to time series analysis, and spectral analysis, of an observed sample $Y(t)$, $t=1, 2,...,T$ suggests that the first step is to identify the memory type (no, short, and long) of the time series.

A no memory time series requires no further analysis; its spectral density is a constant.

A short memory time series is analyzed by representing it (or, more realistically, approximating it) by an ARMA(p,q) scheme:

$$Y(t) + \alpha_p(1) \, Y(t-1)+\ldots+\alpha_p(p) \, Y(t-p)$$
(4)
$$= \epsilon(t) + \beta_q(1) \, \epsilon(t-1)+\ldots+\beta_q(q) \, \epsilon(t-q)$$

where the polynomials

(5)
$$g_p(z) = 1+\alpha_p(1)z+\ldots+\alpha_p(p) \, z^p$$
$$h_p(z) = 1+\beta_q(1)z+\ldots+\beta_q(q) \, z^q$$

are chosen so that all their roots in the complex z-plane are in the region
$\{z:|z|>1\}$ outside the unit circle. Then $g_p(z)$ and $h_q(z)$ are the transfer
functions of invertible filters. $\epsilon(t)$ is assumed to be a white noise time series
which we identify with the innovations $\epsilon(t) = Y^\nu(t)$;

(6) $\qquad \sigma^2_{p,q} = E[\epsilon^2(t)] \neq E[Y^2(t)]$

is an estimator of σ^2_∞. The spectral density of an ARMA(p,q) scheme is

(7) $\qquad f_{p,q}(\omega) = \sigma^2_{p,q} \dfrac{|h_q(e^{2\pi i\omega})|^2}{|g_p(e^{2\pi i\omega})|^2}$

The process of identifying ARMA (p,q) schemes which are adequate (and
parsimonious) approximating models for a time series can be studied rigorously,
and various at least semi-automatic methods are available which are based on
order determining schemes.

A long memory time series Y(·) is usually analyzed by representing it as:
(a) the sum S(·) + N(·) of a long memory signal and a short memory noise, or (b)
as transformable to a short memory time series by a non-invertible filter.
Methods for finding such representations of long memory time series are currently
not as well developed as methods of finding ARMA representations of short
memory time series.

3. Autoregressive Spectral Densities

Our functional inference approach to statistical problems is based on the
premise that _estimation of a function is similar to the problem of model
identification_. A typical problem of statistical model identification seeks to
find the smallest number of significantly non-zero components θ_i of a finite
dimensional parameter $\theta = (\theta_1,\ldots,\theta_k)$. A typical problem of function estimation
seeks to find the smallest finite number of parameters which provide an adequate
("parsimonious") approximation of a function when the function can be
parametrized exactly by a countable infinity of parameters.

The goals of functional inference and model identification are in my view
best pursued simultaneously. One uses finite parameter models, but one
estimates their parameters by methods that can be interpreted as providing either
exact models or approximating models.

Autoregressive spectral estimation is a technique for spectral analysis
based on using autoregressive schemes both as exact models and as approximating
models for stationary time series. An exact model, with an _infinite_ number of
parameters, for a short memory zero mean Gaussian stationary time series Y(t) is
provided by the invertible filters which relate Y(t) and the white noise series
of innovations $Y^\nu(t)$. The AR(∞) representation is

(1) $\qquad Y(t) + \alpha_\infty(1) \, Y(t-1)+\ldots = Y^\nu(t)$;

the MA(∞) representation is

(2) $\qquad Y(t) = Y^\nu(t) + \beta_\infty(1) \, Y^\nu(t-1)+\ldots$

The coefficients $\beta_\infty(k)$ are determined recursively from $\alpha_\infty(j)$ by $\beta_\infty(0) = 1$, and
for k>0

(3) $\qquad \beta_\infty(k) + \sum_{j=1}^{k} \alpha_\infty(j) \, \beta_\infty(k-j) = 0$

The AR(∞) and MA(∞) representations have important implications for spectral

analysis since they provide formulas for the spectral density function $f(\omega)$ alternative to the formula that $f(\omega)$ is the Fourier Transform of $\rho(v)$. One can show that

(4) $\qquad f(\omega) = \sigma_\infty^2 |h_\infty(e^{2\pi i\omega})|^2$,

(5) $\qquad f^{-1}(\omega) = \sigma_\infty^{-2} |g_\infty(e^{2\pi i\omega})|^2$,

defining

(6) $\qquad h_\infty(z) = \sum_{j=0}^{\infty} \beta_\infty(j) z^j$, $g_\infty(z) = \sum_{k=0}^{\infty} \alpha_\infty(k) z^k$.

The AR(∞) transfer function $g_\infty(z)$ can be shown to be the limit of AR(m) transfer functions

(7) $\qquad g_m(z) = 1 + \alpha_m(1) z + \ldots + \alpha_m(m) z^m$

whose coefficients $\alpha_m(j)$ are the coefficients of memory m prediction errors defined as follows:

(8) $\qquad Y^{\mu,m}(t) = E[Y(t)|Y(t-1),\ldots,Y(t-m)]$

(9) $\qquad Y^{\nu,m}(t) = Y(t) - Y^{\mu,m}(t)$

$\qquad\qquad = Y(t) + \alpha_m(1) Y(t-1) + \ldots + \alpha_m(m) Y(t-m)$,

(10) $\qquad \sigma_m^2 = E[|Y^{\nu,m}(t)|^2] \div E[Y^2(t)]$

$\qquad\qquad = E[Y^{\nu,m}(t) Y(t)] \div E[Y^2(t)]$

Explicit formulas for the foregoing predictors, prediction errors, and mean square prediction errors can be obtained from the correlation function $\rho(v)$. A predictor is characterized by the condition that the prediction error is orthogonal (normal) to the predictor variables:

(11) $\qquad E[Y^{\nu,m}(t)Y(t-k)] = 0$, $\quad k=1,\ldots,m$

By substituting (9) into (11) one obtains the famous Yule-Walker equations, defining $\alpha_m(0) = 1$

(12) $\qquad \sum_{j=0}^{m} \alpha_m(j) \rho(j-k) = 0$, $\quad k=1,\ldots,m$

One obtains σ_m^2 by

(13) $\qquad \sigma_m^2 = \sum_{j=0}^{m} \alpha_m(j) \rho(j)$.

For a short memory time series, these equations also hold with $m=\infty$.

The notion of a time series $Y(\cdot)$ being an autoregressive scheme of order p, denoted AR(p), can be defined in terms of prediction theory as follows: $Y(\cdot)$ is an AR(p) if and only if the memory p prediction errors $Y^\nu,P(\cdot)$ is a white noise time series and $\alpha_p(p) \neq 0$. The spectral density of $Y^\nu,P(\cdot)$ can be expressed in terms of $g_p(z)$, the autoregressive transfer function of order p, by

(14) $\qquad f_{Y^\nu,P}(\omega) = \frac{1}{\sigma_p^2} |g_p(e^{2\pi i\omega})|^2 f(\omega)$

If the time series $Y(\cdot)$ is in fact AR(p), then its spectral density equals the function

(15) $\qquad f_p(\omega) = \sigma_p^2 |g_p(e^{2\pi i\omega})|^{-2}$

which we call, in general, the <u>approximating autoregressive spectral density of order p</u>. A time series $Y(\cdot)$ can be regarded as approximated by an AR(p) if

(16) $\qquad \overline{f}_p(\omega) = \frac{f(\omega)}{f_p(\omega)}$

can be regarded as not "significantly" different from a constant. In this way a test of the hypothesis that a time series $Y(\cdot)$ is AR(p) can be converted to a test of the hypothesis that the prediction error time series is white noise.

The sequence of approximating autoregressive spectral densities $f_m(\omega)$, $m=1,2,\ldots$ may be shown to converge as m tends to ∞ to $f(\omega)$ at each ω in $0<\omega\leq 1$ under suitable conditions (see especially Nevai (1979)). Sufficient conditions are that $f(\omega)$ has finite dynamic range (and therefore is bounded above and below) and has a continuous derivative. When an estimator, denoted $\hat{f}_m(\omega)$, of $f_m(\omega)$ is used as an estimator of $f(\omega)$, one has to take into account two kinds of errors, called respectively <u>bias</u> and <u>variance</u>. Bias is a measure of the deterministic difference between $\overline{f_m(\omega)}$ and $\overline{f(\omega)}$, while variance is a measure of the stochastic distance between $\hat{f}_m(\omega)$ and $f_m(\omega)$. As m increases bias decreases while variance increases. This is an example of the fundamental problem of empirical spectral analysis which is how to achieve an optimal balance between bias and variance. When one uses autoregressive spectral estimation, this problem reduces to a question of determining the order m of the approximating autoregressive scheme. The convergence of autoregressive spectral estimators is studied by Berk (1974), Carmichael (1976), and Kromer (1969).

Autoregressive spectral densities can be justified philosophically as the solution to the "maximum entropy" problem: among the spectral densities $f(\omega)$ satisfying

(17) $\quad \int_0^1 e^{2\pi i v \omega} f(\omega)\,d\omega = \rho(v), \quad v=0,\pm 1,\ldots,\pm m$

for specified correlation $\rho(v)$, $v=0,\pm 1,\ldots,\pm m$, the one which maximizes $\int_0^1 \log f(\omega)$ is

(18) $\quad f_m(\omega) = \sigma_m^2 |g_m(e^{2\pi i \omega})|^{-2}$

where $g_m(z)$ is given by (7) in terms of coefficients $\alpha_m(j)$ computed by the Yule-Walker equations (12). This fundamental fact, related to Burg (1968), was proved by VanDenBos (1971). A new proof which avoids the calculus of variations is outlined by Parzen (1981).

The maximum entropy principle provides a motivation or justification for the use of autoregressive spectral estimators. However the maximum entropy principle provides no insight into how to identify an optimal order m, or even what are the effects of different methods of estimating the parameters σ_m, $\alpha_1,\ldots,\alpha_m$. It provides no guidance for how to learn from the data whether the time series is non-stationary (long memory) or stationary (short memory), or whether the best time series model is AR, MA, or ARMA. In my view it is a principle for deriving probability models, rather than statistically fitting models to data.

It should be realized that the maximum entropy principle justifies auto-regressive estimators only for short memory time series (for whom $\log f(\omega)$ is integrable). Autoregressive estimators are justified for long memory time series by the fact that a pure harmonic $Y(t) = A \cos \frac{2\pi}{P} t + B \sin \frac{2\pi}{P} t$ satisfies $Y(t)-\phi Y(t-1)+Y(t-2)=0$ where $\phi = 2 \cos \frac{2\pi}{P}$.

4. An Approach to Empirical Spectral Analysis

The current status of the search for the perfect spectral estimator seems to be as follows: spectral estimation is not a non-parametric procedure, which can be conducted independently of model identification, but is an infinite-parametric procedure. Kay and Marple (1981), in their excellent survey of modern spectral estimation, seem to also come to this conclusion.

The memory type of a time series (long, short, or no memory) and the type of the whitening filter of a short memory time series (AR, MA, or ARMA) must be identified to arrive at the final form of spectral estimator in an applied problem. Approximating autoregressive schemes play two roles: as estimators, and also as diagnostic tools in the process of model identification.

In section 5 we outline the application to model identification of the notions of information divergence, entropy, and cross-entropy (which for Gaussian random variables are expressed in terms of logarithms of mean square prediction errors). The approach described is related to the pioneering work of Akaika(1977).

The basic aim of spectral analysis is to obtain an estimated spectral density with the "right" degree of smoothness in the sense of not introducing spurious spectral peaks, and also resolving nearby spectral peaks. This section outlines our approach to attaining this goal.

A. <u>Pre-processing</u>. To analyze a time series sample $Y(t)$, $t=1,\ldots,T$, one will proceed in stages which often involve the substraction of or elimination of strong effects in order to see more clearly weaker patterns in the time series structure.

The aim of pre-processing is to transform $Y(\cdot)$ to a new time series $\bar{Y}(\cdot)$ which is short memory. Some basic pre-processing operations are memory less transformation (such as square root and logarithm), detrending, "high pass" filtering, and differencing. One usually subtracts out the sample mean

$$\bar{Y} = \frac{1}{T} \sum_{t=1}^{T} Y(t);$$ then the time series actually processed is $Y(t) - \bar{Y}$. If the

mean $\bar{Y}$ is a large number, it should be subtracted; the variations in $Y(t)$ are then the variations of $Y(t)$ about its mean. The sample mean $\bar{Y}$ and sample variance $\hat{R}(0)$ should always be recorded.

B. <u>Sample Fourier Transform by Data Windowing, Extending with Zeroes, and Fast Fourier Transform</u>. Let $Y(t)$ denote a pre-processed time series. The first step in the analysis could be to compute successive autoregressive schemes using domain operations only in the time. An alternative first step is the computation of the sample Fourier transform

(1) $$\tilde{\psi}(\omega) = \sum_{t=1}^{T} Y(t) \exp(-2\pi i \omega t)$$

at an equi-spaced grid of frequencies in $0 \leq \omega \leq 1$, of the form $\omega = \frac{k}{Q}$, $k=0,\ldots,Q-1$. We call Q the spectral computation number. One should always choose $Q \geq T$, and we recommend $Q \geq 2T$.

Prior to computing $\psi(\omega)$, one should extend the length of the time series by addition zeroes to it. Then $\tilde{\psi}(\omega)$, $\omega = \frac{k}{Q}$, can be computed using the Fast Fourier transform.

If the time series may be long memory one should compute in addition a sample "data windowed" Fourier transform

(2) $$\tilde{\psi}_W(\omega) = \sum_{t=1}^{T} Y(t) W(\frac{t}{T}) \exp(-2\pi i \omega t) \quad .$$

To understand the effect of the window, one replaces $Y(t)$ by a spectral representation $Y(t) = \int_0^1 \exp(2\pi i \lambda t) \, d\Psi(\lambda)$; then

$$\tilde{\psi}_W(\omega) = \int_0^1 w_T(\omega - \lambda) \, d\Psi(\lambda) \text{ where } w_T(\lambda) = \sum_{t=1}^{T} W(\frac{t}{T}) \exp(-2\pi i \lambda t).$$

Considerations involved in the choice of data windows are discussed in Harris (1978).

C. <u>Sample Spectral Density</u>. The sample spectral density $f(\omega)$ is obtained essentially by squaring and normalizing the sample Fourier transform;

(3) $$\tilde{f}(\omega) = \frac{|\tilde{\psi}(\omega)|^2}{\frac{1}{Q} \sum_{k=0}^{Q-1} |\tilde{\psi}(\frac{k}{Q})|^2} \quad , \quad \omega = \frac{k}{Q}, \ k = 0, 1, \ldots, Q-1.$$

Over the domain $0 \leq \omega \leq 1$, it integrates to 1.

D. <u>Sample Correlation Function</u>. The sample correlation function

$$\hat{\rho}(v) = \sum_{t=1}^{T-v} Y(t) Y(t+v) \div \sum_{t=1}^{T} Y^2(t) \text{ is computed (using the Fast Fourier}$$

Transform) by

$$\hat{\rho}(v) = \frac{1}{Q} \sum_{k=0}^{Q-1} \exp\left(2\pi i \frac{k}{Q} v\right) \bar{f}\left(\frac{k}{Q}\right) \ ,$$

which holds for $0 \leq v < Q-T = M$. I recommend always taking $M \geq T/3$, and often one should choose $M \geq \bar{T}(2/3)$.

E. Sample Spectral Distribution Function.

$$F(\omega) = 2\int_0^\omega f(\omega')d\omega', \quad 0\leq\omega\leq 0.5 \quad ;$$

the graph of $F(\omega)$ provides qualitative diagnostics of the time series model type.

F. Nonparametric kernel spectral density estimator. An estimator $\hat{f}(\omega)$ of the spectral density is called: parametric when it corresponds to a parametric model for the time series (such as an AR or ARMA model); non-parametric otherwise. A general form of non-parametric estimator is the kernel estimator

$$\hat{f}(\omega) = \sum_{v=-\infty}^{\infty} k\left(\frac{v}{M}\right) \hat{\rho}(v) \ e^{-2\pi i \omega v}, \quad 0\leq\omega\leq1.$$

Two popular choices of kernel are the Parzen window [Parzen (1961)]

$$\begin{aligned} k(t) &= 1-6t^2 + 6t^3 &,& \quad |t| \leq 0.5 &, \\ &= 2\,(1 - |t|)^3 &,& \quad 0.5 < |t| \leq 1 &, \\ &= 0 &,& \quad 1 \leq |t|. \end{aligned}$$

and the spline-equivalent window [Parzen (1958), Cogburn and Davis (1974), Wahba (1980)]

$$k(t) = \frac{1}{1+t^{2r}}$$

where $r \geq 2$ is usually chosen to equal 2 or 4. The problem of determining optimum truncations points M has no general solution; one approach is to choose a large value of M to obtain a preliminary smoothing of the sample spectral density.

G. Autoregressive analysis. The Yule-Walker equations are solved to estimate innovation variances $\hat{\sigma}_m^2$, to which are applied order determining criteria (AIC, CAT) to determine optimal orders $\hat{m}$ and also to test for white noise. The value of $\hat{\sigma}_m^2$ and the dynamic range of the autoregressive spectral estimator $\hat{f}_m(\omega)$ are used to determine the memory type of the time series [Parzen (1982)].

H. Non-stationary autoregression. When a time series is classified as long memory, more accurate estimators of autoregressive coefficients are provided by minimizing a "forward and backward" least squares criterion

$$\sum_{t=m+1}^{T} (Y(t) + \alpha_m(1) Y(t-1)+\ldots\alpha_m(m) Y(t-m))^2$$
$$+ (Y(t-m) +\alpha_m(1) Y(t-m+1) +\ldots+\alpha_m(m) Y(t))^2,$$

or by Burg estimators [for references to descriptions of Burg's algorithm, see Kay and Marple (1981)].

When several harmonics are present in the data, whose frequencies are close together, least squares autoregressive coefficient estimators are more effective than Yule-Walker autoregressive coefficient estimators in providing autoregressive spectral estimators which exhibit the split peaks one would like to see in the estimated spectral density.

I. ARMA analysis. When a time series is classified as short memory the approximating AR scheme is used to form the MA(∞) coefficients which are used to form a subset regression procedure for determining the best fitting ARMA scheme, and the corresponding ARMA spectral density estimator.

J. Inverse correlations and cepstral correlations. Estimators of $\rho i(v)$ and $\gamma(v)$ are computed and used to form nonparametric kernel estimators of $f^{-1}(\omega)$ and log $f(\omega)$, which may provide additional insights into the peaks and troughs to be given significance in the final estimator of the spectrum.

K. Long Memory analysis. In the long memory case, one may want to represent $Y(t)$ as $S(t) + N(t)$, a long memory signal plus short memory noise. An approach to this problem may be provided by treating the sample spectral density values

f(k/Q) as a data batch to be studied by non-parametric data modeling methods using quantile functions [see Parzen (1979)]. The details of such methods are under development.

Examples of the use of our approach to empirical spectral analysis will be provided in the oral presentation. Examples of autoregressive spectral estimators are given by Beamish and Priestley (1981) and Pagano (1980).

5. Time series model identification by estimating information and order determining criteria

Consider two probability densities $f(y)$ and $g(y)$, where $-\infty < y < \infty$. Define: cross-entropy of g with respect to f

(1) $\quad H(f; g) = \int_{-\infty}^{\infty} \{-\log g(y)\} f(y) \, dy;$

entropy of f

(2) $\quad H(f) = H(f; f) = \int_{-\infty}^{\infty} \{-\log f(y)\} f(y) \, dy;$

information divergence of (a model) g from (a true density) f

$\quad I(f; g) = H(f; g) - H(f)$

(3) $\quad = \int_{-\infty}^{\infty} \{-\log \frac{g(y)}{f(y)}\} \, f(y) \, dy.$

Information divergence has some (but not all of the properties of a distance, since $I(f; g) \geq 0$, $I(f; f) = 0$.

The information about a continuous random variable Y in a continuous random variable X is defined by any one of the following equivalent symbols:

$\quad I(Y|X) = I(f_{Y|X}; f_Y),$

(4) $\quad = E_X I(f_{Y|X=x}; f_Y)$

$\quad = \int_{-\infty}^{\infty} dx \, f_X(x) \int_{-\infty}^{\infty} dy \, f_{Y|X=x}(y) \{-\log \frac{f_Y(y)}{f_{Y|X=x}(y)}\}$

Define entropy of Y and conditional entropy of Y given X by

(5) $\quad H(Y) = H(f_Y)$

(6) $\quad H(Y|X) = H(f_{Y|X}) = E_X \, H(f_{Y|X=x})$

One may show that $E_X[H(f_{Y|X=x}; f_Y)] = H(Y)$, which yields the first fundmanetal relation

(7) $\quad I(Y|X) = H(Y) - H(Y|X)$

When X and Y are jointly normal, $H(f_{Y|X=x})$ does not depend on x, and is the entropy of a normal distribution. Therefore for jointly normal random variables X and Y

$\quad H(Y) = \frac{1}{2} \log \Sigma(Y) + \frac{1}{2} (1 + \log 2\pi)$

(8) $\quad H(Y|X) = \frac{1}{2} \log \Sigma(Y|X) + \frac{1}{2} (1 + \log 2\pi)$

$\quad I(Y|X) = -\frac{1}{2} \log \Sigma^{-1}(Y) \Sigma(Y|X)$

We use $\Sigma(Y)$ to denote the variance of Y, and $\Sigma(Y|X)$ to denote the conditional variance of Y given X.

When Y and X are multivariate normal vectors, Σ denotes the covariance matrix, and one can show that

(9) $\quad I(Y|X) = (-\frac{1}{2}) \log \det \Sigma^{-1}(Y) \Sigma(Y|X)$

$\quad = (-\frac{1}{2}) \text{ sum } \log \text{ eigenvalues } \Sigma^{-1}(Y) \Sigma(Y|X)$

The second fundamental relation decomposes information:

(10) $\quad I(Y|X_1, X_2) = I(Y|X_1) + I(Y|X_1; X_1, X_2),$

defining the information about Y in X_2 conditional on X_1

(11) $\quad I(Y|X_1; X_1, X_2) = H(f_{Y|X_1}) - H(f_{Y|X_1, X_2})$

$\quad = H(Y|X_1) - H(Y|X_1, X_2)$

The fundamental approach of information number model identification procedures uses estimators $\hat{I}(Y|X_1; X_1, X_2)$ of information numbers $I(Y|X_1; X_1, X_2)$ as test statistics for discriminating between alternative hypotheses H_0: Y depends on X_1; H_1: Y depends on X_1 and X_2. In other words, the null hypothesis

H_0: X_1 is "sufficient" to predict Y given the information base X_1 and X_2, is equivalent to H_0: $I(Y|X_1; X_1,X_2) = 0$.

Models for a zero mean stationary Gaussian univariate time series $Y(t)$, $t=0$, $\pm 1,\ldots$, can be formulated in terms of information numbers. For $p=1,2,\ldots$, define

(12) $\quad I_p = I(Y|Y_{-1},\ldots,Y_{-p}) = I(Y(t)|Y(t-1),\ldots,Y(t-p))$.

In words, I_p is the information about $Y(t)$ in the p most recent values. Let Y^- denote the infinite past $Y(t-1)$, $Y(t-2),\ldots$. Then $I_\infty = I(Y|Y^-) = \lim_{p\to\infty} I_p$. The hypothesis that the time series $Y(\cdot)$ is an AR(p), autoregressive scheme of order p, is equivalent to

(13) $\quad I(Y|Y_{-1},\ldots Y_{-p}; Y^-) = I_\infty - I_p = 0$

To estimate I_p, one estimates the coefficients of an AR(p). Further

(14) $\quad I_p = -\frac{1}{2} \log \sigma_p^2 \uparrow I_\infty = -\frac{1}{2} \log \sigma_\infty^2 = -\frac{1}{2} \int_0^1 \log f(\omega)\, d\omega$.

Given a time series sample, $Y(t)$, $t=1,2,\ldots T$, of length T, one can calculate successively (using Fast Algorithms such as the Yule-Walker equations) estimators

(15) $\quad \hat{I}_p = -\frac{1}{2} \log \hat{\sigma}_p^2$, $p = 1, 2, \ldots$.

which can be regarded as test statistics for testing white noise, or more precisely AR(0) against AR(p). The work of Akaike (1974, 1977) and Hannan and Quinn (1949) leads one to conjecture that a universal test for white noise (whose theory needs further study) is of the form (for a suitable choice of constant $c \geq 0$, say c=1)

(16) $\quad -\frac{1}{2} \log \hat{\sigma}_p^2 - \frac{p}{T} \log \log T \leq \frac{c}{T}$ for $p=1, 2,\ldots$.

A related conjecture is that optimal orders $\hat{p}$ of approximating autoregressive schemes can be identified by first determining the orders at which are attained the absolute and relative minima of order determining criteria which determine orders p for which $\hat{I}_\infty - \hat{I}_p$ is not significantly different from zero. It should be emphasized that the ultimate decision on the adequacy of a model should be based on a definition of "parsimony of a model" which requires that the spectral distribution function of the residuals (Y| variables in model)$^\nu$ (t) be "parsimoniously" not significantly different from white noise.

Akaike's order determining criterion AIC is defined by

(17) $\quad \text{AIC }(m) = \log \hat{\sigma}_m^2 + \frac{2m}{T}$;

it seeks to determine the order of an exact autoregressive scheme which the time series is assumed to obey. One can raise the objection against it that it does not consistently estimate the order, which is done by a criterion due to Hannan and Quinn (1979):

(18) $\quad \text{AICHQ}(m) = \log \hat{\sigma}_m^2 + \frac{m}{T} \log \log T$

Parzen (1974), (1977) introduced an approximating autoregressive order criterion called CAT (criterion autoregressive transfer function), defined by

(19) $\quad \text{CAT}(m) = \frac{1}{T} \sum_{j=1}^{m} (1-\frac{1}{T}) \hat{\sigma}_j^{-2} - (1- \frac{m}{T}) \hat{\sigma}_m^{-2}$.

In practice CAT and AIC lead in many examples to exactly the same orders. It appears reassuring that quite different conceptual foundations can lead to similar conclusions in practice.

An important application of fitting an approximating autoregressive scheme AR($\hat{p}$) to a time series is the estimation of information numbers which are used to determine the goodness of fit of ARMA (p,q) schemes. An approach to ARMA modeling can be based on estimating $I(Y|Y_{-1},\ldots,Y_{-p}, Y^\nu_{-1},\ldots,Y^\nu_{-q}; Y^-)$.

REFERENCES

Akaike, H. (1970) A fundamental relation between predictor identification and power spectrum estimation. Ann. Inst. Statist. Math. 22, 219-223.

_____. (1974) A new look at the statistical model identification, IEEE Trans. Autom. Contr., AC-19, 716-723.

_____. (1977) On entropy maximization principle, Applications of Statistics, P. R. Krishnaiah, ed., North-Holland, Amsterdam, 27-41.

Beamish, M. and Priestley, M. B. (1981) A study of autoregressive and window spectral estimation. Appl. Statist., 30, 41-58.

Berk, K. N. (1974) Consistent autoregressive spectral estimates. Ann. Statist., 2, 489-502.

Blackman, R. B. and Tukey, J. W. (1959) The Measurement of Power Spectrum from the point of view of Communication Engineering. Dover: New York. (Originally published in Bell Systems Tech. Journal, 37, 185-282 and 485-569 (1958)).

Burg, J. P. (1968) A new analysis technique for time series data, NATO Advanced Study Institute on Signal Processing, Enschede, The Netherlands.

Carmichael, J. P. (1976) The Autoregressive Method, Ph.D. thesis, Statistical Science Division, State University of New York at Buffalo.

_____. (1978) Consistency of the Autoregressive Method of Density Estimation, Technical Report, Statistical Science Division, State University of New York at Buffalo.

Carter, G. C. and Nuttall, A. H. (1982) An analysis of the generalized spectral estimation technique, including a near-optimum method for sonar applications. Proceedings IEEE. To be published.

Chatfield, C. (1979) Inverse autocorrelations, J.R. Statistic. Soc.A 142,363-77.

Childers, D. G. (1978) Modern Spectrum Analysis, New York: IEEE Press.

Cleveland, W. S. (1972) The inverse autocorrelations of a time series and their applications. Technometrics 14, 277-93.

Cogburn, R. and Davis, H. T. (1974) Periodic splines and spectral estimation. Annals of Statistics, 2, 1108-1126.

Durrani, T. S. and Arslanian, A. S. (1982) Windows associated with high resolution spectral estimators. Submitted for publication.

Goodman, H. R. (1957) On the joint estimation of the spectra, cospectrum, and quadrative spectrum of a two-dimension stationary Gaussian process, Ph.D. Thesis, Department of Statistics, Princeton University.

Grenander, U. (1981) Abstract Inference New York: Wiley.

Hannan, E. J. and Quinn, B. G. (1979) The determination of the order of an autoregression, Journal of the Royal Statistical Society, 41, 190-195.

Harris, F. (1978) On the use of windows for harmonic analysis with the discrete Fourier transform, Proc. IEEE, 66, 51-83.

Kay, S. M. and Marple, Stanley L., Jr. (1981) Spectrum Analysis--A Modern Perspective. Proceedings of the IEEE, 69, 1380-1419.

Kromer, R. E. (1969) Asymptotic properties of the autoregressive spectral estimator. Ph.D. Thesis. Statistics Dept., Stanford University.

Munk, W. H. and Snodgress, F. E. (1957) "Measurements of southern swell at Guadalupe Island," Deep-Sea Research, 4, 272-286.

Nevai, Paul G. (1979). An asymptotic formula for the derivatives of orthogonal

polynomials. SIAM J. Math. Anal. 10, 472-477.

Pagano, M. (1980) Some recent advances in autoregressive processes. Directions in Time Series, ed. D. R. Brillinger and G. C. Tiao, Institute of Mathematical Statistics, 280-302 (Comments by H. T. Davis).

Parzen, E. (1957) "On consistent estimates of the spectrum of a stationary time series" Ann. Math. Statist., 28, 329-348.

_____. (1958) "On asymptotically efficient consistent estimates of the spectral density function of a stationary time series" J. Roy. Statist. Soc., B., 20, 303-322.

_____. (1961) "Mathematical considerations in the estimation of spectra" Technometrics, 3, 167-190.

_____. (1964) "An approach to empirical time series analysis" J. Res. Nat. Bur. Standards, Sect. D, 68D, 937-951.

_____. (1967) "Empirical multiple time series analysis" Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, ed. L. LeCam, and J. Neyman, University of California Press, Vol. I, pp. 305-340.

_____. (1967) "The role of spectral analysis in time series analysis" Review of the International Statistical Institute, 35, 125-141.

_____. (1968) "Statistical spectral analysis (single channel case) in 1968" Proceedings of NATO Advanced Study Institute on Signal Processing, Enschede, Netherlands.

_____. (1969) "Multiple time series modeling" Multivariate Analysis-II, ed. P. Krishnaiah, Academic Press: New York, 389-409.

_____. (1970) "Statistical inference on time series by RKHS methods, II" Proceedings 12th Seminar Canadian Mathematical Congress, ed. R. Pyke, Canadian Mathematical Congress, Montreal 1-37.

_____. (1974) "Some Recent Advances in Time Series Modeling" IEEE Transactions on Automatic Control, AC-19, 723-730.

_____. (1977) "Multiple Time Series: Determining the Order of Approximating Autoregressive Schemes" Multivariate Analysis-IV, Ed. P. Krishnaiah, North Holland: Amsterdam, 283-295.

_____. (1979) "Nonparametric Statistical Data Modeling" Journal of the American Statistical Association, (with discussion), 74, 105-131.

_____. (1979) "A Density-Quantile Function Perspective on Robust Estimation" Robustness in Statistics, ed. R. Launer and G. Wilkinson, New York: Academic Press, 237-258.

_____. (1980) "Time Series Modeling, Spectral Analysis, and Forecasting" Directions in Time Series Analysis, ed. D. R. Brillinger and G. C. Tiao, Institute of Mathematical Statistics.

_____. (1981) Autoregressive Spectral Estimation, Log Spectral Smoothing, and Entropy. IEEE Spectral Estimation Workshop (Hamilton, Canada).

_____. (1982) "Time Series Model Identification and Prediction Variance Horizon," Applied Time Series Analysis II, ed. D. Findley, New York: Academic Press, 415-447.

_____. (1983) Autoregressive Spectral Estimation, Handbook of Statistics III, ed. D. Brillinger and P. Krishnaiah, Amsterdam: North Holland.

VanDenBos, A. (1971) Alternative interpretation of maximum entropy spectral analysis, IEEE Trans. Inform. Theory, IT-17, 493-494.

Wahba, Grace (1980). Automatic smoothing of the log periodogram, Journal of the American Statistical Assn., 75, 122-132.

Wiener, N. (1930) Generalized harmonic analysis. Acta. Math., 55, 117-258.